


[Subscribe \(Full Service\)](#) [Register \(Limited Service, Free\)](#) [Login](#)

 Search: ☒ The ACM Digital Library ☐ The Guide



THE ACM DIGITAL LIBRARY



searching similar documents jaccard distance and percentage

 Terms used: [searching similar documents](#) [jaccard distance](#) [percentage](#)

 Sort results by 
☒ [Save results to a Binder](#)

 Refine these results w  
 Try this search in The

 Display results 
☐ [Open results in a new window](#)

Results 1 - 19 of 19

### 1 [Temporal profiles of queries](#)



Rosie Jones, Fernando Diaz

July 2007 ACM Transactions on Information Systems (TOIS), Volume 25 Issue 3

Publisher: ACM

 Full text available: [pdf\(430.31 KB\)](#)

 Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 12, Downloads (12 Months): 316, Citation Count: 1

Documents with timestamps, such as email and news, can be placed along a timeline. The timeline for a set of documents returned in response to a query gives an indication of how documents relevant to that query are distributed in time. Examining the ...

Keywords: Time, ambiguity, event detection, language models, precision prediction, query classification, temporal profiles

### 2 [Exploiting correlated keywords to improve approximate information filtering](#)



Christian Zimmer, Christos Tryfonopoulos, Gerhard Weikum

July 2008 SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval

Publisher: ACM

 Full text available: [pdf\(510.74 KB\)](#)

 Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 61, Downloads (12 Months): 56, Citation Count: 0

Information filtering, also referred to as publish/subscribe, complements one-time searching since users are able to subscribe to information sources and be notified whenever new documents of interest are published. In approximate information filtering ...

Keywords: Peer-to-Peer (P2P), approximate publish/subscribe, distinct-value (DV) estimation, distributed information filtering (IF), information systems

### 3 [Robust and efficient fuzzy match for online data cleaning](#)



Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, Rajeev Motwani

June 2003 SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data

Publisher: ACM

 Full text available: [pdf\(271.47 KB\)](#)

 Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index term](#)

Bibliometrics: Downloads (6 Weeks): 22, Downloads (12 Months): 309, Citation Count: 48

To ensure high data quality, data warehouses must validate and cleanse incoming data tuples from external sources. In many situations, clean tuples must match acceptable tuples in *reference tables*. For example, product name and description fields ...

4 [Localized signature table: fast similarity search on transaction data](#)

❖ Qiang Jing, Rui Yang, Panos Kalnis, Anthony K. H. Tung

November 2004 CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management

Publisher: ACM

Full text available:  [pdf\(200.77 KB\)](#)Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 5, Downloads (12 Months): 45, Citation Count: 0

Recently, techniques for supporting efficient similarity search over huge transaction datasets have emerged as an important research area. Several indexing schemes have been proposed towards this direction. Typically, these schemes provide a tradeoff ...

Keywords: data mining, indexing, similarity search, transaction data

5 [Detection of Duplicate Defect Reports Using Natural Language Processing](#)

Per Runeson, Magnus Alexandersson, Oskar Nyholm

May 2007 ICSSE '07: Proceedings of the 29th international conference on Software Engineering

Publisher: IEEE Computer Society

Full text available:  [pdf\(268.53 KB\)](#)Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 2, Downloads (12 Months): 225, Citation Count: 2

Defect reports are generated from various testing and development activities in software engineering. Sometimes two reports are submitted that describe the same problem, leading to duplicate reports. These reports are mostly written in structured natural ...

6 [Distributed group-based cooperative caching in a mobile broadcast environment](#)

❖ Chi-Yin Chow, Hong Va Leong, Alvin T. S. Chan

May 2005 MDM '05: Proceedings of the 6th international conference on Mobile data management

Publisher: ACM

Full text available:  [pdf\(333.27 KB\)](#)Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index term](#)

Bibliometrics: Downloads (6 Weeks): 15, Downloads (12 Months): 103, Citation Count: 5

Caching is a key technique for improving data retrieval performance of mobile clients. The emergence of state-of-the-art peer-to-peer communication technologies now brings to reality what we call "cooperative caching" in which mobile clients not only ...

Keywords: cooperative caching, mobile computing, mobile data broadcast, peer groups, peer-to-peer computing

7 [Communications of the ACM: Volume 51 Issue 1](#)

❖ January 2008 issue Volume 51 Issue 1

Publisher: ACM



Full text available:  [pdf\(5.97 MB\)](#)  [digital edition](#) Additional Information: [full citation](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 553, Downloads (12 Months): 3458, Citation Count: 0

8 [Communications of the ACM: Volume 51 Issue 2](#)

❖ February 2008 issue Volume 51 Issue 2

Publisher: ACM

Full text available:  [pdf\(3.89 MB\)](#)  [digital edition](#) Additional Information: [full citation](#)

Bibliometrics: Downloads (6 Weeks): 321, Downloads (12 Months): 1983, Citation Count: 0

9 [A survey of Web metrics](#)

Devanshu Dhyani, Wee Keong Ng, Sourav S. Bhowmick

December 2002 ACM Computing Surveys (CSUR), Volume 34 Issue 4

Publisher: ACM

Full text available: pdf(269.28 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index term](#)

Bibliometrics: Downloads (6 Weeks): 70, Downloads (12 Months): 707, Citation Count: 18

The unabated growth and increasing significance of the World Wide Web has resulted in a flurry research activity to improve its capacity for serving information more effectively. But at the heart of these efforts lie implicit assumptions about "quality" ...

Keywords: Information theoretic, PageRank, Web graph, Web metrics, Web page similarity, query metrics

10 [THESUS: Organizing Web document collections based on link semantics](#)

Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis, Michalis Vazirgiannis

November 2003 The VLDB Journal — The International Journal on Very Large Data Bases, Volume 12 Issue 4

Publisher: Springer-Verlag New York, Inc.

Full text available: pdf(262.85 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index term](#)

Bibliometrics: Downloads (6 Weeks): 9, Downloads (12 Months): 90, Citation Count: 5

The requirements for effective search and management of the WWW are stronger than ever. Currently Web documents are classified based on their content not taking into account the fact that these documents are connected to each other by links. We claim ...

Keywords: Document clustering, Link analysis, Link management, Semantics, Similarity measures, World Wide Web

11 [Supporting Intelligent Web search](#)

Maurice Coyle, Barry Smyth

October 2007 ACM Transactions on Internet Technology (TOIT), Volume 7 Issue 4

Publisher: ACM

Full text available: pdf(1.63 MB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 44, Downloads (12 Months): 719, Citation Count: 1

Search engines continue to struggle to provide everyday users with a service capable of delivering focussed results that are relevant to their information needs. Moreover, traditional search engines really only provide users with a starting point for ...

Keywords: Collaborative search, Web search, explanation, interaction history, personalization, visualisation

12 [Automatic thesaurus construction](#)

Dongqiang Yang, David M. Powers

January 2008 ACSC '08: Proceedings of the thirty-first Australasian conference on Computer science - Volume 74, Volume 74

Publisher: Australian Computer Society, Inc.

Full text available: pdf(213.31 KB)

Additional Information: [full citation](#), [abstract](#), [references](#)

Bibliometrics: Downloads (6 Weeks): 4, Downloads (12 Months): 11, Citation Count: 0

In this paper we introduce a novel method of automating thesauri using syntactically constrained distributional similarity. With respect to syntactically conditioned cooccurrences, most popular approaches to automatic thesaurus construction simply ignore ...


Keywords: distribution, similarity, syntactic dependency

13 [Building structured web community portals: a top-down, compositional, and incremental approach](#)

Pedro DeRose, Warren Shen, Fei Chen, AnHai Doan, Raghu Ramakrishnan

September 2007 VLDB '07: Proceedings of the 33rd international conference on Very large data bases

Publisher: VLDB Endowment

Full text available:  pdf(342.90 KB)

Additional Information: [full citation](#), [abstract](#), [references](#)

Bibliometrics: Downloads (6 Weeks): 10, Downloads (12 Months): 143, Citation Count: 4

Structured community portals extract and integrate information from raw Web pages to present unified view of entities and relationships in the community. In this paper we argue that to build portals, a top-down, compositional, and incremental ...


14 [Efficient similarity joins for near duplicate detection](#)



Chuan Xiao, Wei Wang, Xuemin Lin, Jeffrey Xu Yu

April 2008 WWW '08: Proceeding of the 17th international conference on World Wide Web

Publisher: ACM

Full text available:  pdf(327.62 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 33, Downloads (12 Months): 118, Citation Count: 0

With the increasing amount of data and the need to integrate data from multiple data sources, a challenging issue is to find near duplicate records efficiently. In this paper, we focus on efficient algorithms to find pairs of records such that their ...

Keyw ords: near duplicate detection, similarity join

15 [Instance-based attribute identification in database integration](#)

Cecil Eng H. Chua, Roger H. L. Chiang, Ee-Peng Lim

October 2003 The VLDB Journal — The International Journal on Very Large Data Bases, vol. 12 Issue 3

Publisher: Springer-Verlag New York, Inc.

Full text available:  pdf(220.13 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 7, Downloads (12 Months): 118, Citation Count: 0

Abstract. Most research on attribute identification in database integration has focused on integrating attributes using schema and summary information derived from the attribute values. No research has attempted to fully explore the use of attribute values ...

Keyw ords: Attribute identification, Database integration, Measures of association


16 [SlideSeer: a digital library of aligned document and presentation pairs](#)



Min-Yen Kan

June 2007 JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries

Publisher: ACM

Full text available:  pdf(1.33 MB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 9, Downloads (12 Months): 175, Citation Count: 0

Research findings are often transmitted both as written documents and narrated slide presentations. As these two forms of media contain both unique and replicated information, it is useful to combine and align these two views to create a single synchronized ...

Keyw ords: SlideSeer, digital library, fine-grained alignment, presentations (slides), synchronized media

17 [Organizing and searching the world wide web of facts -- step two: harnessing the wisdom of the crowds](#)



Marius Paşca

May 2007 WWW '07: Proceedings of the 16th international conference on World Wide Web

Publisher: ACM

Full text available: pdf(204.15 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 18, Downloads (12 Months): 291, Citation Count: 1

As part of a large effort to acquire large repositories of facts from unstructured text on the Web, seed-based framework for textual information extraction allows for weakly supervised extractor class attributes (e.g., side effects and generic ...

Keywords: class attributes, fact extraction, knowledge acquisition, named entities, unstructured text, web search queries

# 18 [Mining approximate functional dependencies and concept similarities to answer imprecise queries](#)



Ullas Nambiar, Subbarao Kambhampati

June 2004 WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases colocated with ACM SIGMOD/PODS 2004

Publisher: ACM

Full text available: pdf(195.43 KB)

Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#)

Bibliometrics: Downloads (6 Weeks): 1, Downloads (12 Months): 30, Citation Count: 3

Current approaches for answering queries with imprecise constraints require users to provide distance metrics and importance measures for attributes of interest. In this paper we focus on providing a domain and end-user independent solution for supporting ...

Keywords: approximate functional dependencies, imprecise queries, tuple similarity

# 19 [Learning Social Networks from Web Documents Using Support Vector Classifiers](#)

Masoud Makrehchi, Mohamed S. Kamel

December 2006 WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence

Publisher: IEEE Computer Society

Full text available: pdf(833.00 KB)

Additional Information: [full citation](#), [abstract](#), [index terms](#)

Bibliometrics: Downloads (6 Weeks): 1, Downloads (12 Months): 221, Citation Count: 0

Automatic generation of a social network requires extracting pair-wise relations of the individual this research, Learning social network from incomplete relationship data is proposed. It is assumed that only a small subset of relations between ...

## Results 1 - 19 of 19

The ACM Portal is published by the Association for Computing Machinery. Copyright © 2008 ACM,  
[Terms of Usage](#) [Privacy Policy](#) [Code of Ethics](#) [Contact Us](#)

Useful downloads: [Adobe Acrobat](#) [QuickTime](#) [Windows Media Player](#) [Real Player](#)